融入情境强度的客户行为模式挖掘及变化侦测①

琚春华[1,2], 帅朝谦[1]
(1. 浙江工商大学计算机与信息工程学院, 杭州 310018; 2. 浙江工商大学现代商贸研究中心, 杭州 310018)

摘要: 现有的客户行为分析往往忽略了客户群体的情境信息,使得行为模式及其变化分析存在局限性. 文章首先定义了客户的情境、情境强度和行为变化,并对其进行了量化处理; 其次,提出了融入情境强度约束的行为模式挖掘方法和模式变化侦测方法,并进一步提取了造成行为变化的关键情境. 文中对情境强度的考虑,使得规则变化的敏感度加大,改进了海量数据稀疏下关联规则支持度、置信度和敏感度低的缺点. 实验和分析证明了方法的可行性和有效性.

关键词: 情境强度; 客户行为; 约束频繁模式; 变化侦测; 推荐策略
中图分类号: TP311.13 **文献标识码:** A **文章编号:** 1007-9807(2014)08-0060-14

0 引 言

电子商务市场竞争日趋激烈,为了更精确了解用户兴趣爱好并提供令其满意的个性化服务,促使信息系统必须主动认知、获取、分析客户个体信息[1]. 客户的消费行为不仅受到需求、产品本身的影响,还很大程度上取决于客户的个体情境,如居民消费价格指数(CPI)、教育背景、收入水平、天气等. 融入情境能够更加准确地挖掘客户行为,而情境的变化能够引起消费行为的变化,挖掘客户的行为变化能够使企业决策者及时了解消费者需求及兴趣变化.

目前有关消费者行为分析的研究相对较多,其中较具代表性成果有: 张紫琼等[2]对在线评论情感分析的研究现状与进展动态进行了归纳和分析,重点论述了现有研究采用的主要方法和关键技术. Liu 等[3]运用决策树对网络消费者的行为变化进行了挖掘. 但由于决策树是基于分类的方法,不能实现预测顾客全部的行为变化模式,使得采用决策树对顾客行为变化的挖掘不能达到预期效果. Song 等[4]分析消费者在网上购物行为集合的关联规则、挖掘变化的行为,采用规则匹配的方法侦测各种类型的行为变化,通过显著的规则变化来评价变化程度和兴趣度. 但往往因为数据稀疏而变化敏感度低. 吴斌和马超[5]从海量旅行数据中挖掘旅客类型和环境因素之间内在的、隐含的相关性,提出基于关系延展路径约束的关联规则并行挖掘算法,其中的相关性没有考虑情境强度. 常亚平等[6]研究了虚拟社区知识共享对消费者行为的影响,并提出了关系强度、社区活跃度等 5 个维度表达信息发送者的能力,用以测试影响消费者购买行为的关键因素,其不足之处在于对特殊情境(如教育背景、地理位置因素等)的分析还有待加强. Yun[7]考虑了项的重要度,提出长度递减支持约束加权频繁模式挖掘算法; Lee 等[8]考虑项在不同标准下的重要度,在最小支持和最小置信度两个约束的基础上根据项的重要度动态地改变约束条件在项上的取值; Liu 等[9]只考虑不统一的最小支持度方法,但如果项目的支持小于阈值,这个项是不值得考虑的,而且仍然将约束条件局限于规则的兴趣度,没有考虑规则来源的情境背景. 上述对消费者行为模式的研究,部分考

① 收稿日期: 2012-06-21; 修订日期: 2013-09-15.
基金项目: 国家自然科学基金资助项目(71071141); 国家科技支撑计划子课题资助项目(2012BAI34B01-5); 浙江省自然科学基金资助项目(LY14F020002).
作者简介: 琚春华(1962—),男,浙江常山人,博士,教授. Email: juchunhua@hotmail.com

虑了情境信息,部分考虑了属性的重要度,但没有将情境强度融入到行为模式挖掘的兴趣度度量中.

挖掘在情境知识背景下的行为模式及模式变化,实际上就是将情境属性加入到记录集中作为模式. 挖掘的约束条件,情境强度又能作为兴趣度的度量方法,在海量高维数据下避免支持度、置信度的低敏感度. 融入情境强度的行为模式挖掘框架如图 1 所示. 本文的主要工作包括: 1) 从客户的情境信息出发,量化定义客户的情境、情境强度和行为变化; 2) 将情境作为项加入到频繁模式挖掘过程中,在情境约束下缩小建构最大频繁模式树的搜索空间; 3) 在规则结果集的基础上,比较不同情境下规则的置信度,确定情境强度; 4) 提出融入情境强度约束行为模式变化侦测方法,提取造成行为变化的关键情境.

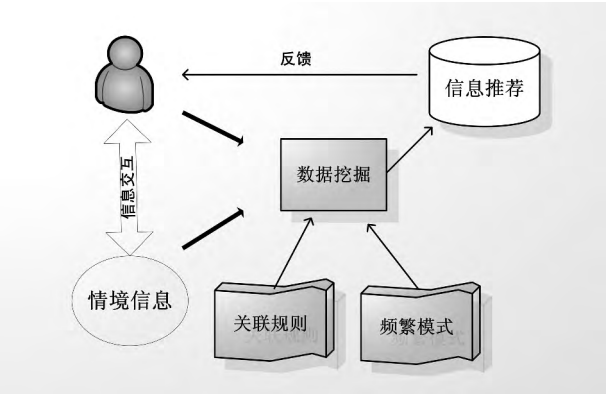

图 1　融入情境强度的行为模式挖掘框架

Fig. 1 Behavior pattern mining framework with context intensity constraints

1　情境与情境强度

相关领域的研究成果包括: 1) 基于人口统计学特征的用户行为研究,主要考虑用户的静态属性(统计学特征)对动态行为特征的影响; 2) 基于关联规则的用户兴趣模式,并考虑了权重和支持度变化; 3) 行为模式变化侦测. 以上研究存在的共同缺陷在于在用户行为模式挖掘中尚未融入隐含概念漂移② 特征的、包括静态和动态属性的情境知识,使得规则模式的准确性和识别变化的敏感度存在不足.

1.1　相关工作

1) 基于背景知识的推荐　Bosnjak 等[10] 运用 Fishbein 模型对网络用户的态度进行了标准化度量,采用 scheffe 检验等方法探索了影响在线用户消费态度的相关因素、态度之间的关系以及影响因素,有效地证明了由用户背景组成的客户特征对消费态度和购物决策有重要影响. Hong 等[11] 的情境感知系统研究实践指出,除了位置是非常重要的判断服务依据,用户 profile 文件(如性别,年龄,风格,偏好等)对推荐服务准确度也有重要影响. Ding 和 Li[12] 设计与开发的推荐系统讨论了随着时间推移而变化的不同时间加权的协同过滤问题. Muhammad 等[13] 提出了基于半自动多元属性特征模式挖掘方法.

2) 规则模式变化侦测　规则模式挖掘方面,现有的研究成果大多数是基于约束的关联规则挖掘. Stefano 和 Francesco[14] 从提高挖掘性能、获得真正用户感兴趣的模式的角度开展了研究. 赵旭俊等[15] 展示了先验知识指导 FP-Tree 的构造方法,该方法能减少 FP-Tree 中大量与挖掘兴趣无关的冗余信息,但该文献没有考虑情境因素和强度. Yun 和 Ryu[16] 考虑在噪声环境下权重和支持度微小变化会严重影响结果集的情况,提出了挖掘近似加权频繁模式的概念. 在基于权重的模式挖掘框架基础上,融入近似因子来放宽权重支持和最小支持度平衡要求,虽然引入了权重使得挖掘结果更有效,但是没有考虑用户的情境属性及权重. 宋国杰等[17] 研究了数据流环境中异常模式的提取与趋势监测,提出了基于强度比率的异常模式挖掘度量标准,并利用最小二乘法回归分析进行异常模式趋势监测,实时有效地度量和提取异常模式,该文献对情境分析后异常模式度量和提取很有借鉴意义.

3) 多维关联规则　涉及两个或多个维或谓词的关联规则称为多维关联规则,不重复谓词的多维关联规则称作维间关联规则,某些谓词多次出现的称作混合维关联规则. 搜索频繁项集变成了搜索频繁谓词集,Kamber 等[18] 在挖掘之前使用预定义的概念分层对量化属性离散化,数值属性的值用区间替代. 如果任务相关的结果数据存

② 概念漂移特征即指人们的兴趣度会随着时间的变化发生改变,用户在 t_1 时刻的购买行为与其在 t_2 时刻的购买行为是有差异的.

放在关系表中，改进的 Apriori 算法 k 或 $k+1$ 次表扫描就能找出所有频繁谓词，而不是频繁项集. Ananthanarayana 等[19]提出了基于聚类的多维关联规则挖掘方法，该方法能在复杂情境环境下挖掘有效的语义规则. Li 等[20]基于多维情境提出了基于扩展的 priori 的关联规则范式，并通过真实情境的仿真与分析验证了该规则的有效性与合理性. Kweku 和 Osei[21]提出基于肯定和否定空间关联规则的情境模式挖掘算法.

上述研究成果分别从情境属性，约束条件下的规则模式挖掘及模式推荐策略等方面开展了研究，是本文情境强度下的规则模式推荐的基础和方法论. 下文将以网购行为为例做客户行为分析.

1.2 情境

客户的网购行为取决于其个体背景特征和购物环境，也就是情境背景. 本文考虑的客户情境，是信息系统从会员信息、用户注册信息、浏览点击记录、人口统计学特征以及交易发生时容易获得的动态情境知识. 情境特征决定个体的网购行为习惯，关键的特征情境决定了购物的特殊购物行为模式和行为的变化趋势. 融入情境的行为模式研究，即在情境背景的约束条件下，获取与客户相关的行为模式，并从情境强度及变化角度来分析行为和行为变化.

定义 1 情境(context) 客户的情境由外部自然环境、社会环境、个体特征等组成，具体包括时间、季节、温度、教育水平、收入水平、CPI、评价、满意度等; 从客户维度上，使用元组 $Context(cid, date, C_0)$ 表示客户 cid 在 date 时间的情境特征，C_0 是客户 cid 在 date 时间的全局情境集合，含情境属性集和兴趣度 Interestingness.

定义 2 客户行为变化(online shopping behavior changing) 客户行为变化是指网购个体或群体在一定的时间范围内，在线客户行为在商品类目、购买频度、消费金额、搜索点击等方面的变化. 这种变化受其情境影响，可以以规则的形式描述.

1.3 情境强度模型与算法

情境的说明与描述 情境由形如 $C_o\{c_{11}c_{21}c_{34}c_{43}\cdots, b_o \times e_0\}$ 的集合来表示，包含情境属性或者情境属性序列($c_{11}c_{21}c_{34}c_{43}\cdots$，可以含 n 个属性，当 $n=1$ 时为单个情境，当 $n>1$ 时为多个情境组成的情境属性序列)，相关度、强度组成的情境兴趣度等，其中相关变量说明如下:

C_o: 客户对象 O 的全局情境;

P: 条件属性，用于约束情境属性;

c_{ij}: 情境属性 c_i 可扩展到不同的划分 c_{ij}，代表情境属性的不同取值，如 c_{i1}: 男性，c_{i2}: 女性;

b_o: 客户对象 O 的初始情境属性，由初始客户行为数据测定;

b_{ij}: 客户对象 O 与第 i 个情境属性第 j 种取值 c_{ij} 的相关程度;

e_o: 客户对象 O 的初始情境强度，由初始数据给定;

e_{ij}: 对象 O 关于第 i 个情境属性第 j 种取值的强度;

n: 相关的情境属性个数.

定义 3 情境强度(context intensity) 不同的情境属性、情境属性集合对购物行为及其变化有不同程度的影响，通过强度来度量某个情境或情境集合对行为模式的影响程度.

通常有关情境强度描述如下，一个客户 cid 在 date 时间的情境 $C_{context}$ 或情境集合 C_o，首先对其进行量化表示，强度值通过情境强度算法来确定.

定义 4 关联度 为准确反映情境 $C_{context}$(这里 $C_{context}$ 即为 c_{ij}，$\overline{C}$ 即 $c_{i\bar{j}}$)对用户行为 Q 的贡献率，这里引用文献[20]有关关联度的定义

$$correlation(C_{context} \Rightarrow Q) = \frac{p(Q \mid C_{context})}{p(Q \mid \overline{C_{context}})} \quad (1)$$

此外，需要注意的是有时需要考虑排除某种情境外的关联度问题，此时，若原情境为 $C_{context}$，则以上情况可用符号 $\overline{C}$ 或 $c_{i\bar{j}}$ 表示，代表含义为非情境 c_{ij} 属性下的取值.

关联度的说明: 关联度可以很好地描述情境与行为的相关程度. 关联度等于 1，说明 $C_{context}$ 的存在与否和 Q 无关; 关联度大于 1，则说明 $C_{context}$ 的出现对 Q 有诱导作用，关联度越大，相关程度越高. 此外，本文引入“初始化相关度确定策略”以及“更新策略”，其主要目的是通过客户行为数据的挖掘找出与情境信息的相关度. 其中，初始化

关联度确定策略为: 大量历史信息确定,没有历史信息时,捕获一定量的流数据进行初始化,存在冷启动的问题; 而更新策略: 较大粒度时间间隔下统计(count) ,进行阶段性更新. 这两种策略用以描述客户行为受到包括情境等多种信息的影响而产生的变化趋势.

本文将情境强度作为一个影响因素引入到关联度计算公式中,以此提高该公式的有效性.

情境强度可以对单个情境进行计算,也可对情境属性集进行计算,其中每一个情境(集合) 的强度对应一个具体规则. 根据不同角度,可将情境进行划分,诸如"年龄区间"的人口统计学情境和"页面点击喜好"的行为情境; 或分为静态情境和动态情境(如性别男是静态的,气温升高是动态的) ; 或分为一般情境和叠加情境(如高温和持续高温) . 支持度变大说明最近的大众(群体为单位) 行为趋向,置信度变大说明最近的群体内部的行为趋向. 情境强度有如下求解算法.

算法 1　基于属性分拆的情境强度算法

算法思想: 通过比较多个情境在结果部分相同的情况下规则置信度的差异,求得情境强度. 这里,情境强度 e_{ij} 的度量是综合考虑情境属性的相关程度和规则的置信度,并通过函数来描述情境强度.

步骤 1　对于规则

$$P \Rightarrow Q(\sup port = S, Confidence = C_{\text{context}})$$

将条件 P 分解为 $p_1 + p_2 + \cdots + p_n$,其中

$$p_1 \Rightarrow Q(\sup port = S_1, Confidence = C_1),$$

$$p_2 \Rightarrow Q(\sup port = S_2, Confidence = C_2),$$

$$\cdots\cdots,$$

$$p_n \Rightarrow Q(\sup port = S_n, Confidence = C_n)$$

均是

$$P \Rightarrow Q(\sup port = S, Confidence = C)$$

的子规则,且 S_1、S_2、$\cdots$、S_n 均是在其规则下的统计量.

步骤 2　计算情境强度 e_{ij},它由关联度 b_{ij} 与情境 c_{ij} 下的置信度决定,即关联度与情境越高 e_{ij} 越大,基于此函数可定义为

$$e_{ij} = f(b_{ij}, c_{ij}) = b_{ij} \frac{(C_{\text{context}} + c_{ij})}{C_{\text{context}}} \tag{2}$$

算法 2　一种基于属性集合的情境强度算法

算法思路: 如果情境 C_{context} 下,规则 $C_{\text{context}} \Rightarrow Q$ 成立,则需要满足条件

$$p(Q \mid C_{\text{context}}) > p(Q \mid \overline{C_{\text{context}}})$$

即在 C_{context} 情境约束下用户行为 Q 发生的概率,大于 $\overline{C_{\text{context}}}$ 情境约束下用户行为 Q 发生的概率. 即 C_{context} 情境约束下,Q 行为得到提升度(关联度)

$$correlation(C_{\text{context}} \Rightarrow Q) = \frac{p(Q \mid C_{\text{context}})}{p(Q \mid \overline{C_{\text{context}}})} > 1$$

进一步可得到一个度量

$$\mu = p(Q \mid C_{\text{context}}) - p(Q \mid \overline{C_{\text{context}}})$$

$\mu > 0$,是规则 $C_{\text{context}} \Rightarrow Q$ 成立的必要条件; $\mu = 0$,情境 C_{context} 与行为 Q 相互独立; $\mu < 0$,是"情境 C_{context} 约束下不倾向于有行为 Q 的发生"规则成立的必要条件. 当 $\mu > 0$ 且满足某个显著的阈值时,是规则 $C_{\text{context}} \Rightarrow Q$ 成立的充分条件.

步骤 1　计算关联度

$$correlation(C_{\text{context}} \cap P \Rightarrow Q) = \frac{p(Q \mid C_{\text{context}})}{p(Q \mid \overline{C_{\text{context}}})}$$

得

$$\mu = p(Q \mid C_{\text{context}} \cap P) - p(Q \mid \overline{C_{\text{context}}} \cap P)$$

其中 C_{context} 是情境属性,P 是条件属性,Q 是结果属性.

步骤 2　计算情境属性 C 的强度

$$e(C_{\text{context}} \cap P \Rightarrow Q) = \frac{\mu}{\sqrt{\dfrac{p(Q \mid P \cap \overline{C})\,[1 - p(Q \mid P \cap \overline{C})]}{N}}} \tag{3}$$

b_{ij} 和 e_{ij} 的乘积 $b_{ij}\ e_{ij}$ 称为兴趣度(interestingness) ,即某个用户对象对某一情境的相关程度与该对象对于该情境强度的乘积,表明客户基于情境的强度和关联度,对所产生的行为的兴趣度,它是可累加的.

2　情境强度约束的行为模式挖掘

本文通过分析不同时刻的数据集,在进行关联规则挖掘的同时,融入交易记录原始客户的情境信息,并反映在规则集中,从而体现客户情境强

度对行为变化的影响. 主要的解决思路和方法如下: 1) 比较不同时刻的规则集,符合相应的规则匹配阈值的记录集合形成含情境属性信息的变化规则集; 2) 计算情境强度,并通过变化程度的度量提取变化规则集; 3) 根据情境强度变化与客户行为变化的关系,从而更新推荐策略. 变化侦测和推荐策略更新框架如图 2 所示,其详细描述如下: 通过关联规则挖掘"T 时刻数据集"得到下方的"T 时刻规则集",而非"$T+K$ 时刻规则集". 首先采用关联规则挖掘出不同时刻数据集的规则集,即 T 时刻和 $T+K$ 时刻数据集的规则集; 其次比较 T 时刻规则集和 $T+K$ 时刻规则集,把符合规则匹配阈值的规则集形成含情境属性信息的变化规则集; 随后对该变化规则集分别进行变化程度度量和情境强度计算得到显著变化规则集和情境强度集,通过对两者的分析比较,更新推荐策略,最终得到推荐策略库.

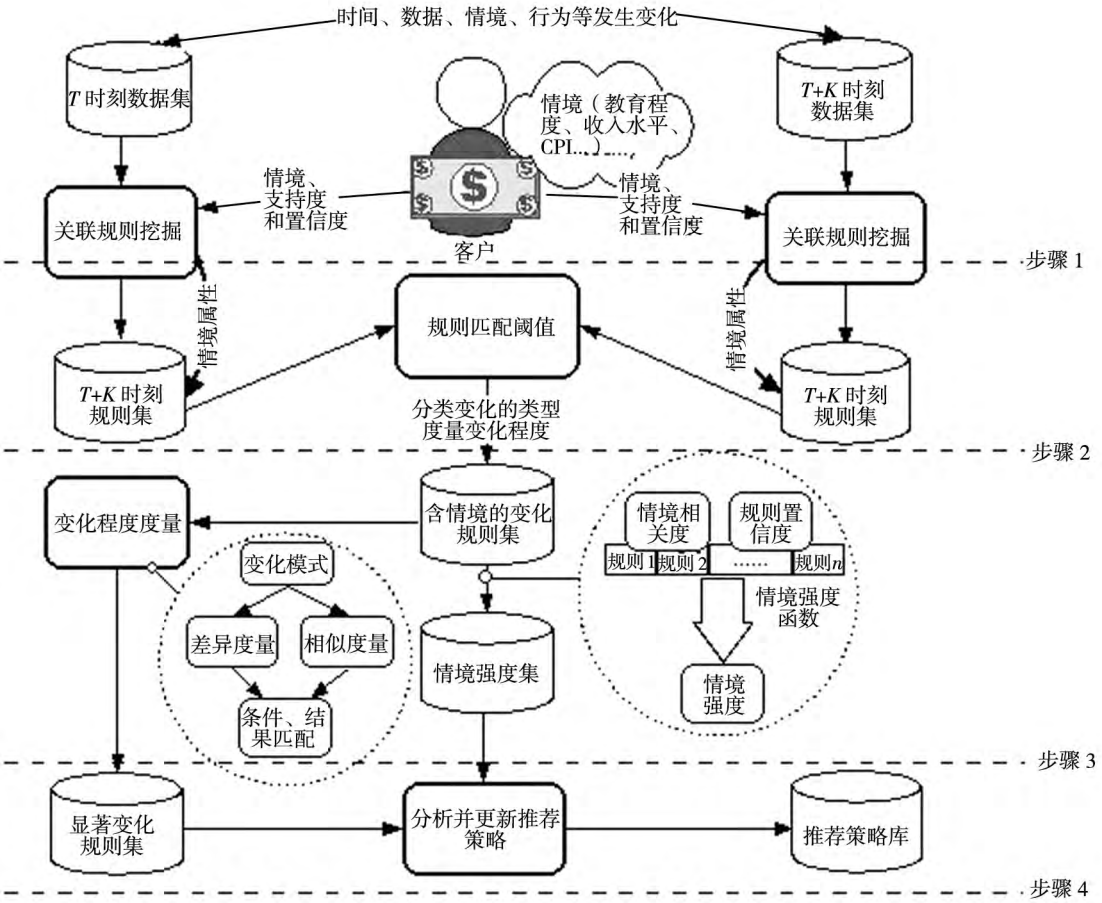

图 2　融入情境强度约束的模式变化侦测与推荐策略更新模型

Fig. 2 Model of pattern change detection and recommended strategy updating with context intensity

对于一个训练集(包含客户情境的 N 条原始交易记录 T),给定支持度阈值 $\lambda_{\text{sup port}}$ 和置信度阈值 $\lambda_{\text{confidence}}$,通过计算情境强度、模式规则的推荐强度来进行行为度量.

算法 3　行为度量策略算法

步骤 1　挖掘剔除了情境信息的交易记录,提取满足支持度阈值 $\lambda_{\text{sup port}}$ 和置信度阈值 $\lambda_{\text{confidence}}$ 的关联规则模式,计算每条规则的支持度和置信度;

步骤 2　抽取情境因素(如天气、CPI、高/中收入人群、学历),用 c_{ij} 表示,其中 i 表示第 i 个情境因素,j 表示某个情境 c_i 的第 j 个划分或簇;

步骤 3　计算情境强度 e_{ij} 和相关度 b_{ij},以及 $b_{ij}e_{ij}$;

步骤 4　根据全局情境,得出不同模式(如 $P \Rightarrow Q$ 的规则)在当前情境划分 c_{ij}(或 c_{ij} 的集合 C_k,即条件 P 中包含情境 c_{ij} 或 C_k)下,$P \Rightarrow Q$ 的支持度和置信度;

步骤 5　比较不同划分或 C_k(或 c_{ij})下规则 $P \Rightarrow Q$ 的支持度 $\lambda'_{\text{sup port}}$ 和置信度 $\lambda'_{\text{confidence}}$,其值越高表明情境与规则的相关度越高,进而确定推荐强度(strength of recommendations)

$$SoR = \frac{\lambda'_{\text{confidence}}}{\lambda_{\text{confidence}}}$$

3　情境强度约束的行为变化侦测方法

本文第 2 节给出了客户情境强度对行为变化的影响的 3 个步骤,其中第 3 点根据情境强度变化与客户行为变化的关系,从而更新推荐策略是其关键. 为了详细阐述该步骤,本节首先定义行为模式的变化及变化的形式,接着阐述行为模式变化的度量方法,然后提出行为模式变化规则提取算法,最后展示其变化原因的追溯从而进行更新推荐策略.

从情境到行为的关联规则角度来讲,情境变化导致网购行为的变化,即意味着规则的变化. 变化的形式包括: 1) 规则保持; 2) 新兴模式; 3) 异常(条件/结果)变化; 4) 新生/灭亡的规则. 规则变化的分类遵循了 Dong 和 Li[23],Song 等[24],以及 Chen 等[25]的研究成果.

3.1　行为模式变化及变化度量

D_i^t, D_j^{t+k}: 在 t,$t+k$ 时刻的数据集;

R_i^t,R_j^{t+k}: 在 t,$t+k$ 时刻挖掘得到的隐含情境属性的关联规则集;

r_i^t, r_j^{t+k}: 规则集合 R_i^t,R_j^{t+k} 中的相应规则,其中 $i = 1,2,\cdots, |R^t|$, $j = 1,2,\cdots, |R^{t+k}|$;

$Sup^t(r_i)$: t 时刻 r_i 的支持度.

定义 5 新兴模式　包含情境属性的条件和行为属性的结果部分没有随时间变化,但支持度显著变化的模式. 即对于规则 r_j^{t+k},如果满足以下两个条件: 一是 r_i^t, r_j^{t+k} 的条件和结果部分没有随时间变化; 二是支持度 support 显著变化了,则称其为 r_i^t 的新兴模式.

定义 6 异常(条件/结果)变化　两条规则的条件部分没有随时间变化,但结果部分有显著差异,即行为发生变化. 即如果 r_i^t, r_j^{t+k} 的条件部分是相似的,而两条规则的结论部分存在显著差异,则 r_j^{t+k} 是关于 r_i^t 的异常变化;

定义 7 新生/灭亡模式　条件和结果都不在原有的集合中,或者规则的条件和结果都在集合中消失了. 即如果所有的条件和结论都与 R^t 中的 r_i^t 有显著差异,则 r_j^{t+k} 是增加的规则,如果所有的条件和结论都与和 R_j^{t+k} 中的 r_j^{t+k} 有显著差异,则 r_i^t 是消失的规则. 行为变化度量由规则匹配阈值 RMT 控制,其中人为定义的一个值,可根据具体的情况进行动态调整,也可根据历史数据预先定义,如图 3.

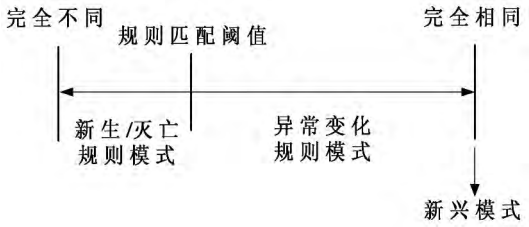

图 3　规则的不同类型变化

Fig. 3 Different typical changes of regulars

3.2　行为模式变化规则提取算法

采用规则匹配方法侦测不同类型的变化规则. 在 t,$t+k$ 时刻挖掘得到的关联规则集,结合用户指定的 RMT,采用算法 4 来进行行为模式变化规则提取(包括规则保持、新兴模式、异常(条件/结果)变化、新生/灭亡等 4 条).

算法 4　行为模式变化规则提取算法

步骤 1　计算时刻 t,$t+k$ 的每一条规则的最大相似度值(r_i^t 和 r_j^{t+k} 的相似度是规则相似度[25]度量);

步骤 2　对每一条规则 r_i^t,计算 r_i^t 和 r_j^{t+k} 之间的差异度量值(r_i^t 和 r_j^{t+k} 的差异度是规则差异度度量);

步骤 3　使用最大相似度值和差异度量值来进行不同变化规则的分类.

定义 8 规则相似度

$$s_{ij} = \begin{cases} \dfrac{l_{ij}\sum\limits_{k\in A_{ij}} x_{ijk}c_{ij}y_{ij}}{|A_{ij}|}, & \text{若 } |A_{ij}| \neq 0 \\ 0, & \text{否则} \end{cases} \quad (4)$$

定义 9　规则差异度

$$\delta_{ij}=\begin{cases}\dfrac{l_{ij}\sum\limits_{k\in A_{ij}}x_{ijk}}{|A_{ij}|}-y_{ij}, & \text{若}\ |A_{ij}|\neq 0, c_{ij}=1\\ -y_{ij}, & \text{否则}\end{cases} \quad (5)$$

其中: $l_{ij}=\dfrac{r_i^t \text{和} r_j^{t+k}\ \text{条件部分相同的属性数量}}{r_i^t \text{和} r_j^{t+k}\ \text{条件部分的属性数量的较大值}}$ 为条件部分的属性匹配度; c_{ij} 为结果部分的属性匹配度,相同为 1、不相同为 0; x_{ijk} 和 y_{ij} 为条件部分和结果部分的每一个属性的值匹配度,相同为 1,否则为 0; $|A_{ij}|$ 即为 r_i^t 和 r_j^{t+k} 条件部分相同的属性数量; $\dfrac{l_{ij}\sum\limits_{k\in A_{ij}}x_{ijk}}{|A_{ij}|}$ 是 r_i^t 和 r_j^{t+k} 的条件部分的相似度; $c_{ij}y_{ij}$ 是结果部分的相似度. r_i^t 的最大相似度值 $s_i=\max(s_{i1},s_{i2},\cdots,s_{i|R_{i+k}|})$; r_j^{t+k} 的最大相似度值 $s_j=\max(s_{1j},s_{2j},\cdots,s_{|R_i|j})$. 如果 $s_i<RMT$,那么 r_i^t 被称为灭亡模式; 如果 $s_j<RMT$,那么规则 r_j^{t+k} 称为新生模式.

4 实验仿真及分析

4.1 问题背景

以表 1 所示网购行为为例,首先在条件属性为空的条件下,对情境属性和行为属性进行关联度分析. 根据定义 4,关联度定义为

$$correlatin(C\Rightarrow Q)=\frac{p(Q|C)}{p(Q|\bar{C})}$$

表 1 情境属性与行为属性

Table 1 Context attributes and behavior attributes

序号	情境属性	数量	行为属性	数量
1	已婚有孩子	45	母婴产品	18
2	已婚无孩子	10	母婴产品	3
3	未婚	55	母婴产品	4
4	男	40	男装	28
5	女	60	男装	12
…	…	…	…	…

“已婚有孩子”与购买“母婴产品”的关联度为 (18/45) / [(3+4)/(10+55)] = 3.71;

“男性”与购买“男装”的关联度为(28/40) / (12/60) =3.5;

“女性”与购买“男装”的关联度为 2/7.

当情境背景属性是(0,1)分布时,即取值只有两种情况,记为取值 a 和 b,则 $a=>Q$ 的相关度与 $b=>Q$ 互为倒数.

然后分两种情形进行情境强度的分析.

1) 情境属性在得到关联规则集之后再添加到规则的条件部分

以表 2 为例,从中不难看出: 在不同情境下(如 2、3、4、5)关联规则的支持度和置信度是有差异的,这说明不同的情境下规则的兴趣度不同; 甚至,同一情境下不同的取值(如 2、3)造成规则兴趣度的不同,这说明情境属性的取值对规则兴趣度是受强度影响的.

表 2 情境属性对关联规则兴趣度的影响

Table 2 Effect of context attributes on interestingness of association rules

序号	融入情境属性	情境	支持度(%)	置信度(%)
1	$\{I_1,I_2\}\ \{I_3\}$	Ø	5	60
2	$\{(c_{11});\ I_1,I_2\}\ \{I_3\}$	c_{11}	4	75
3	$\{(c_{12});\ I_1,I_2\}$	c_{12}	1	50
4	$\{I_3\}\{(C_{11},C_{21},C_{31});\ I_1,I_2\}\ \{I_3\}$	C_1	3.5	80
5	$\{(C_{12},C_{22},C_{32});\ I_1,I_2\}\ \{I_3\}$	C_2	1	45
…	…	…	…	…

2) 情境属性直接融入频繁模式和关联规则集生成过程

以表 3 为例,按照算法 1 和算法 2,有两种方案.

表 3 融入情境强度的关联规则

Table 3 Association rules with context intensity

序号	融入情境属性	情境	支持度	置信度
1	$\{I_1,I_2\}\ \{I_3\}$	Ø	5%	80%
2	$\{C_{11},I_1,I_2\}\ \{I_3\}$	$\{C_{11}\}$	4.2%	90%
3	$\{C_{22},I_1,I_2\}\ \{I_3\}$	$\{C_{22}\}$	1.5%	30%
4	$\{C_{11},C_{21},C_{31},I_1,I_2\}\ \{I_3\}$	$\{C_{11},C_{21},C_{31}\}$	3.2%	68%
5	$\{C_{11},C_{22},C_{31},I_1,I_2\}\ \{I_3\}$	$\{C_{11},C_{22},C_{31}\}$	1.0%	22%
6	$\{C_{21},I_1,I_2\}\ \{I_3\}$	$\{C_{21}\}$	3.5%	75%
7	$\{C_{31},I_1,I_2\}\ \{I_3\}$	$\{C_{31}\}$	4.0%	85%
8	$\{C_{11},C_{31},I_1,I_2\}\ \{I_3\}$	$\{C_{11},C_{31}\}$	3.8%	82%
…	…	…	…	…

方案一:

步骤 1 如上述比较的情形 2) 中表 3 所示;

步骤 2 抽取情境因素如上述表 3 的第 3 列所示;

步骤 3 根据公式

$$e_{ij} = f(b_{ij}, c_{ij}) = b_{ij}\frac{(C + c_{ij})}{C}$$

假设任意 C 的取值都只有两种情况,有

$$e_{11} = f(b_{11}, c_{11}) = b_{11}\frac{(C + c_{11})}{C} = 5.25 \times \frac{80\% + 90\%}{80\%} = 11.16$$

同理可得其他情境属性的强度 $e_{21} = f(b_{21}, c_{21}) = b_{21}\frac{(C + c_{21})}{C} = \frac{7}{3}\frac{80\% + 75\%}{80\%} = 4.52$,

$$e_{22} = f(b_{22}, c_{22}) = b_{22}\frac{(C + c_{22})}{C} = \frac{3}{7}\frac{80\% + 30\%}{80\%} = 0.59,$$

$$e_{31} = f(b_{31}, c_{31}) = b_{31}\frac{(C + c_{31})}{C} = 4\frac{80\% + 85\%}{80\%} = 8.25,$$

……

步骤 4　对于$\{C_{11}, C_{22}, C_{31}\}$,有 $A_0(\{C_{11}, C_{22}, C_{31}\}) = b_{11}e_{11} + b_{22}e_{22} + b_{31}e_{31} = 22.25$

对于$\{C_{11}, C_{21}, C_{31}\}$,有 $A_0(\{C_{11}, C_{21}, C_{31}\}) = b_{11}e_{11} + b_{21}e_{21} + b_{31}e_{31} = 23.93$

步骤 5　对步骤 4 的值进行归一化到区间[0,1],得到情境强度约束的关键关联规则及推荐强度 SoR.

方案二:

在规则 $\{C_{11}, C_{21}, C_{31}, I_1, I_2\} \Rightarrow \{I_3\}$ 下,情境属性及其条件部分、结果部分分别是

$$C_0 = \{c_{11}, c_{21}, c_{31}\}, P = \{I_1, I_2\},$$

$$Q = \{I_3\}, n = 1\,000$$

则有

$$e(C \cap P \Rightarrow Q) = \frac{p(Q \mid C \cap P) - p(Q \mid P \cap \overline{C})}{\sqrt{\frac{p(Q \mid P \cap \overline{C})(1 - p(Q \mid P \cap \overline{C}))}{N}}} = \frac{3.2\% - 1.8\%}{\sqrt{\frac{1.8\%(1 - 1.8\%)}{1\,000}}} = 3.33$$

同理可从其他规则中得到购买$\{I_3\}$时各情境属性的强度值. 由于强度值的不同,在不同情境下的推荐强度也随之变化.

4.2　数据集的搜集与处理

实验数据集来源于某电子商务网站,从该网站 2011 — 01 ~ 2012 — 01 的交易数据中随机抽取 10 万条记录,主要目的是挖掘分析情境强度约束下用户类目偏好模式. 首先对数据集进行预处理,主要对属性缺失值和异常值进行处理,并对处理后的数据进行分类标记. 经过系统获取信息和分类打标,数据集中包含的情境属性有: CPI、天气、季节、节日、性别、年龄、教育水平、收入水平、地理位置等. 其中 CPI(<0; [0%,1%); [1%,3%); ≥3),天气(持续高温; 持续低温; 升温; 降温; 舒适),季节(春; 夏; 秋; 冬),节假日(非节假日; 节假日; 重要节假日),性别(男; 女),婚姻(年轻单身; 年轻已婚无孩子; 年长已婚有孩子),教育水平(高中及以下; 本科及以下; 研究生及以上),收入水平(低; 中; 高),地区(乡村; 中小城市; 一线城市),商品类目偏好(女装; 男装; 手机、数码; 鞋、箱包; 运动、户外; 化妆品). 具体的数据导入结果如表 4 所示.

附表 1 展示了包含情境知识的购物篮数据片段. 从上述数据集中,随机抽取 80% 的数据,在最小项集为 2,支持度阈值 $\lambda_{support}$ 为 0.06,置信度阈值 $\lambda_{confidence}$ 为 20% 的条件下,采用改进的 Apriori 算法挖掘出的规则有 180 条,表 5 是部分规则. 从这 180 条规则中过滤出形如实例演示中的规则,继而通过算法 1 或算法 2 获得该规则成立时所有交易记录的情境强度,以{ C_ID,情境属性,情境强度} 的形式存储为先验知识,为后续的推荐提供"强度"支持.

表 4　数据源变量展示表

Table 4 List of data resource variables

属性	类别	取值范围	数据类型
CPI	分类	CPI: >3	字串
天气	分类	升温	字串
季节	分类	冬、夏等	字串
节假日	分类	节假日	字串
性别	分类	女,男	字串
婚姻	分类	年轻单身等	字串
教育	分类	本科及以下等	字串
收入	分类	低,中,高	字串
地区	分类	乡村等	字串

表 5 情境知识的数据关联规则集($T1$ 时刻,训练集,10/180)

Table 5 Data association rule set of context knowledge (period $T1$, training set, 10/180)

序号	规则	支持度(%)	自信度(%)
1	CPT [1%,3%),中小城市,女 = > 女装	11.825	30.73
2	已婚有孩子,女 = > 女装	11.577	48.07
3	一线大城市,本科及以下,男装 = > 鞋、箱包	7.980	29.53
4	升温,男 = > 男装,运动、户外	12.678	20.43
5	CPT [1%,3%),重要节假日,男 = > 手机、数码	15.602	33.29
6	CPT [1%,3%),非节假日,男 = > 手机、数码	10.563	23.03
7	非节假日,一线大城市,女,运动、户外 = > 鞋、箱包	10.180	27.93
8	舒适,春,男,已婚有孩子,母婴产品 = > 男装	8.643	62.50
9	男,年轻单身,研究生及以上 = > 男装,手机、数码	7.500	47.68
10	女,年轻单身 = > 女装	19.870	64.58

变化的规则集如表 6 由 STATE 列给出,在所展示的 10 条规则中,规则 1、2、6 是新生模式,3、4、5、7、10 是新兴模式,其变化趋势表现在支持度、置信度的变化(变大或变小) ; 规则 8、9 是异常模式,以规则 8 为例,相对于表 2 的规则 4,前者在规则的结果部分发生了显著变化: 从“运动、户外”变化为“家用电器”,是典型的异常模式,而推动模式变化的关键情境因素是季节(夏季) .

对于基于规则的推荐系统,根据基于用户个体情境强度的规则模式挖掘和变化模式提取,实时更新推荐规则.

4.3 基于情境强度的客户行为挖掘及侦测算法评测与分析

本文的研究成果在上述实例分析和数据集分析中体现了两个方面,一是数据流中概念漂移特征提取和特征模式变化的挖掘,二是融入造成用户兴趣行为发生变化的情境背景信息使得能够更准确地挖掘用户的行为特征模式和模式变化.

表 6 情境知识的数据关联规则集($T2$ 时刻,测试集,10/167)

Table 6 Data association rule set of context knowledge(period $T2$, training set, 10/167)

序号	规则	状态	支持度(%)	自信度(%)	变化
1	CPT [3%, + ∞),一线大城市,女 = > 运动、户外	新生模式	9.277	33.33	条件部分变化
2	CPT [3%, + ∞),中小城市,女 = > 女装	新生模式	11.009	32.68	条件部分变化
3	一线大城市,女 = > 运动、户外	新兴模式	11.005	39.88	支持度上升,置信度上升
4	女,年轻单身 = > 女装	新兴模式	21.079	65.47	支持度上升,置信度上升
5	一线大城市,本科及以下,男装 = > 鞋、箱包	新兴模式	7.980	29.53	支持度下降,置信度下降
6	CPT [3%, + ∞),重要节假日,男 = > 家用电器	新生模式	7.563	40.00	条件部分变化
7	年轻单身,重要节假日,男 = > 手机、数码	新兴模式	16.667	37.30	支持度上升,置信度上升
8	升温,男 = > 男装,家用电器	异常模式	8.563	23.03	结果部分发生变化
9	男,已婚有孩子,母婴产品 = > 家用电器	异常模式	13.675	42.50	结果部分发生变化
10	男,已婚有孩子,母婴产品 = > 男装	新兴模式	11.640	63.80	支持度上升,置信度上升

本节将展示概念漂移特征模式挖掘算法和融入情境背景信息的行为模式挖掘算法的挖掘结果,应用到某电子商务个性化推荐系统中的情况,以此作评测和分析. 该电子商务个性化推荐系统与其搜索系统紧密结合,推荐系统的推荐结果展示在搜索结果页的右侧,推荐结果的生成必须依靠获取用户的信息(历史交易、搜索、点击、浏览、收藏类目偏好等) 和当前的搜索行为(搜索词、类目、时间等) ,是类协同过滤推荐系统.

首先,分析由搜索日志系统识别的情境属性和行为数据对推荐系统提供的有效信息. 搜索日志系统识别的情境属性包括用户搜索的时间(time) 、历史的搜索点击偏好(常用搜索关键词(c_query) ,代表高价、低价、平均价等的价格偏好(pp) 等) 、当前网站推广活动(activity) 、及用户信息(用户(id) 、性别(gender) 、年龄(age) 、类目购买力(cpp) 等) ,行为数据则包括用户的不同时间搜索使用的关键词(query) 、点击商品的类目(category) 、点击商品的价格(price) 、点击量(click_num) 等.

取某年 8 月 31 日 10 时 ~ 15 时和后一天 9 月 1 日 10 时 ~ 15 时这两个不同时段 t,$t + k$ 的用户搜索点击流数据 D_i^t、D_j^{t+k} ,首先根据漂移特征选择算法,识别了不同用户在不同情境下对商品价格的兴趣变化,如表 7 所示.

表 7　融入情境信息的用户搜索点击流概念漂移挖掘结果

Table 7 Click-stream concept-drifting mining result of user searching with context information

数据集	算法	规则特征	结论	算法效果
D_i^t、D_j^{t+k}	漂移特征选择算法	概念漂移特征模式：time：夏末 + activity：推广活动 = > query 连衣裙 + price：低（点击均价降低）；pp：高价 + click_num：高 = > price：高价（价格分布变化，出现第二个波峰）	夏末，商家推出折扣活动导致连衣裙搜索点击均价降低；高消费能力的点击量上升，且使得点击的价格分布出现概念漂移	识别了造成价格偏好漂移的原因，对“pp：高价”的用户推荐高价优质商品

然后根据情境强度约束的行为模式挖掘模型和行为模式变化规则提取算法，对用户搜索点击流数据 D_i^t、D_j^{t+k} 进行挖掘，得到关键情境属性和变化的规则，如表 8 所示.

此外，对于上述算法的挖掘效果，给出两方面的评测数据. 首先对用户搜索点击流数据 D_i^t、D_j^{t+k} 的在“query：连衣裙”下的用户点击商品的价格分布作估计. 对于数据量大点击流，D_i^t、D_j^{t+k} 采用 EM 算法作高斯参数估计，由于用户搜索点击流概念漂移挖掘结果已经表明，大量的低购买力用户点击低价商品且高购买力用户点击高价商品的量也在上升，造成“query：连衣裙”的商品点击价格很可能趋向双高斯分析. 价格取 $\log2(price+1)$，通过 EM 算法拟合得到高斯曲线分别如图 4 和图 5，其中纵坐标为概率，横坐标为价格 $price$ 取 $\log2(price+1)$. 对比两图可知，在 t 和 $t+k$ 两个时刻，用户群特征在商品价格点击这个维度被区分得更明显，总体点击商品均价在降低，但价格分布发生变化，出现第二个波峰，是典型的概念漂移现象. 但是仅仅通过统计学方法分析价格点击分析，如图 4 和图 5，不能得出发生概念漂移的用户群和发生概念漂移的原因，本文提出的方法则能较好地解决此问题.

表 8　融入情境信息用户搜索点击流行为模式及变化挖掘结果

Table 8 Click-stream behavior pattern and change mining result of user searching with context information

数据集	算法	规则特征	结论	算法效果
D_i^t、D_j^{t+k}	情境强度算法、行为模式变化规则提取算法	新兴规则模式 1：time：夏末 = > query：连衣裙 + price：低价 新兴规则模式 2：pp：高价 + click_num：高 = > price：高价 新生模式 1：time：夏末 + activity：推广活动 = > query：连衣裙 + price：低价； 新生模式 2：pp：中价 + activity：推广活动 = > query：连衣裙 + price：高价；	新兴规则模式 1 的规则支持度由 45.3% 提升到 53.7%；新兴规则模式 2 说明拥有高价购买力的用户搜索点击量上升且该规则支持度置信度均上升；新生模式 1 和新生模式 2 是新出现的规则模式，加入到规则模式库中；“pp：中价”用户在“activity：推广活动”期间搜索时推荐高价的连衣裙的强度提升；“activity：推广活动”是关键情境属性	识别了造成不同用户价格偏好，提升挖掘新兴规则模式，提升相应情境下规则模式推荐强度

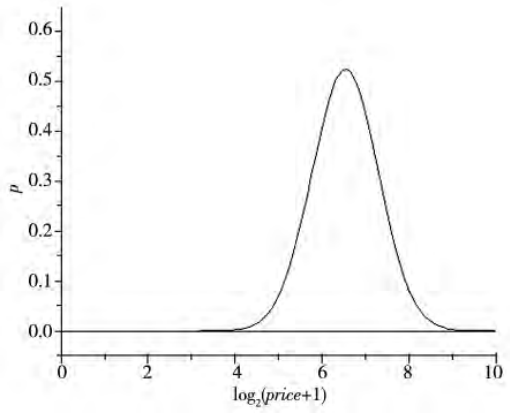

图 4　t 时刻用户搜索点击价格分布

Fig. 4 Price distribution of user searching click at period t

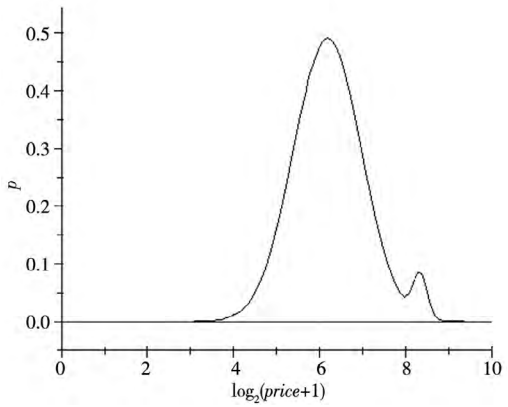

图 5　$t+k$ 时刻用户搜索点击价格分布

Fig. 5 Price distribution of user searching click at period $t+k$

其次,以推荐效果作为评测标准. 选取 50 个带有概念漂移特征(这里仅考虑价格维度的兴趣漂移) 的用户添加到实验数据中,随后抽样选取无概念漂移用户添加到实验数据集中进行对比测试. 主要对比指标是推荐系统命中率和购买转化率,前者是指推荐列表中的商品链接被用户点击的百分比,后者则是用户通过推荐链接进入商品详情页面浏览并且该用户当天发生该商品成交记录的百分比.

表 9 为推荐系统的命中率和购买转化率对比结果,由此表可知: 无论是推荐系统的命中率还是购买转化率,加入本文算法的结果都得到明显的提升; 此外,无论样本的大小发生何种变化,该算法总能提升以上两个指标,说明该算法具有较好的鲁棒性.

表 9 推荐系统的命中率和购买转化率对比

Table 9 Comparison of hit rate and purchase percent conversion of recommended system

用户数		推荐系统命中率(%)		购买转化率(%)	
概念漂移用户数	无概念漂移用户数	未加入本文算法	加入本文算法	未加入本文算法	加入本文算法
50	0	0.91	1.71	0.00	0.20
50	40	3.95	4.78	0.14	0.25
50	100	5.23	5.59	0.21	0.27
50	180	5.93	6.24	0.28	0.34
50	320	6.28	6.39	0.39	0.43
50	365	6.47	6.52	0.41	0.44
50	400	6.49	7.01	0.43	0.45

图 6、7(横坐标为无概念漂移用户数,概念漂移用户数均为 50,纵坐标是百分比) 绘制的算法对比关系图也进一步反映了算法在求解结果方面的优越性.

目前,越来越多的电商企业使用推荐策略服务于用户. 好的推荐系统能够帮助用户更好地找到所需的商品、应用和服务等,能够提高商家的销售额,创造更大的利润. 推荐系统命中率和购买转化率是各大电商企业最为关注的两个指标,本算法的作用在于可以有效的提升这两个指标,其重要性不言而喻.

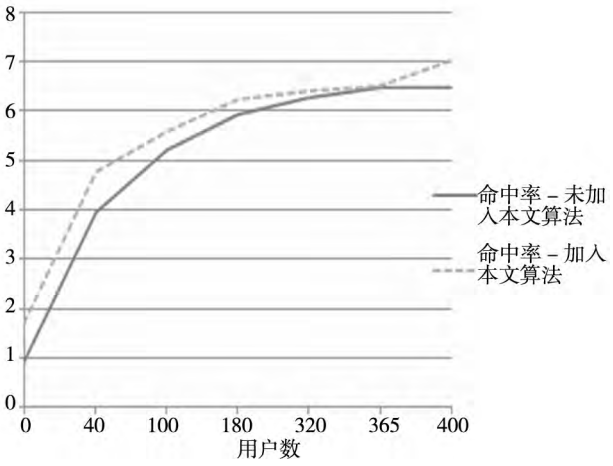

图 6 命中率对比

Fig. 6 Comparison of hit rate

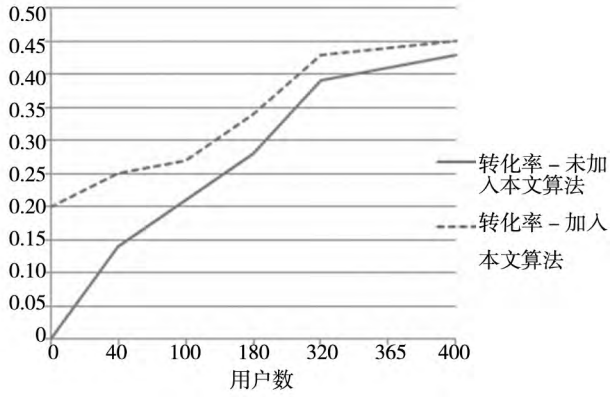

图 7 购买转化率

Fig. 7 Purchase percent conversion rate

5 结束语

通过分析加入客户情境的关联规则,规则模式中加入情境属性使得兴趣度(如置信度、支持度) 发生变化,说明不同的情境下规则的兴趣度不同. 而同一情境属性取不同值时,兴趣度也不同,说明情境属性的取值对规则兴趣度是有强烈影响的. 文中基于情境强度的定义,挖掘不同时刻交易记录的规则集,抽取情境强度,并通过变化程度的度量形成显著变化规则集. 根据情境强度集合和显著变化规则集合,得到情境强度变化与客户行为变化的关系,从而更新推荐策略.

通过实验可知,融入情境信息的行为模式挖

掘，不仅能够减小确定模式时的搜索空间，还能挖掘对影响消费行为情境的强度，从而更准确地挖掘情境约束下的行为模式. 此外，情境背景强度约束下的数据流挖掘及变化侦测方法中，如何确定恰当有效的情境属性，加强时间窗口特征，从而准确反映规则模式是急需解决的问题. 在研究行为模式变化时，仅考虑融入概念漂移情境，如何识别随机噪声并有效地区分概念漂移和随机噪声，以及在大数据量情况下保证算法的有效性等是另一个需要深入研究的问题. 最后，若考虑“情境”项目之间的自相关性，对客户行为模式挖掘的影响是今后又一个需要研究的问题.

参 考 文 献:

[1] 王　茜，杨莉云，杨德礼. 面向用户偏好的属性值评分分布协同过滤算法 [J]. 系统工程学报，2010，25(4)：561－568.
Wang Qian，Yang Liyun，Yang Deli. Collaborative filtering algorithm based on rating distribution of attributes faced user preference [J]. Journal of Systems Engineering，2010，25(4)：561－568. (in Chinese)

[2] 张紫琼，叶　强，李一军. 互联网商品评论情感分析研究综述 [J]. 管理科学学报，2010，13(6)：84－96.
Zhang Ziqiong，Ye Qiang，Li Yijun. Literature review on sentiment analysis of online product reviews [J]. Journal of Management Sciences in China，2010，13(6)：84－96. (in Chinese)

[3] Liu B，Hsu W，Han H S，et al. Mining changes for real-life applications [C] // Second International Conference on Data Warehousing and Knowledge Discovery，2000：337－346.

[4] Song H S，Kim J K，Kim S H. Mining the change of customer behavior in an internet shopping mall [J]. Expert Systems with Applications，2001，21(3)：157－168.

[5] 吴　斌，马　超. 一种旅行数据约束关联规则挖掘算法 [J]. 计算机工程与应用，2010，46(20)：129－132.
Wu Bin，Ma Chao. Constrained association rule mining algorithm for travel data [J]. Computer Engineering and Applications，2010，46(20)：129－132. (in Chinese)

[6] 常亚平，刘兴菊，阎　俊，等. 虚拟社区知识共享之于消费者购买意向的研究 [J]. 管理科学学报，2011，14(4)：86－96.
Chang Yaping，Liu Xingju，Yan Jun，et al. Research on knowledge sharing in virtual communities on consumer purchase intention [J]. Journal of Management Sciences in China，2011，14(4)：86－96. (in Chinese)

[7] Yun Unil. An efficient mining of weighted frequent patterns with length decreasing support constraints [J]. Knowledge-Based Systems，2008，21(8)：741－752.

[8] Lee Yeong-Chyi，Hong Tzung-Pei，Lin Wen-Yang. Mining association rules with multiple minimum supports using maximum constraints [J]. International Journal of Approximate Reasoning，2005，40(1/2)：44－54.

[9] Liu B，Hsu W，Ma Y. Mining association rules with multiple minimum supports [C] // International Conference on Knowledge Discovery and Data Mining，1999：337－341.

[10] Bosnjak M，Galesic M，Tuten T. Personality determinants of online shopping：Explaining online purchase intentions using a hierarchical approach [J]. Journal of Business Research，2007，60(6)：597－605.

[11] Hong Jong-yi，Suh Eui-ho，Kim Sung-Jin. Context-aware systems：A literature review and classification [J]. Expert Systems with Applications，2009，36(4)：8509－8522.

[12] Ding Y，Li X. Time weight collaborative filtering [C] // CIKM 05：Proc. of the 14th ACM int. conf. on Information and knowledge management，2005：485－492.

[13] Muhammad S，Muhammad S，Aziz G. Context based positive and negative spatio-temporal association rule mining [J]. Knowledge-Based Systems，2013，37(1)：261－273.

[14] Stefano B，Francesco B. Extending the soft constraint based mining paradigm [C] // Lecture Notes in Computer Science (Knowledge Discovery in Inductive Databases)，2007，4747：24－41.

[15] 赵旭俊，张继福，蔡江辉. 约束频繁模式树及其构造方法研究 [J]. 小型微型计算机系统，2010，31(4)：682－685.

Zhao Xujun, Zhang Jifu, Cai Jianghui. Constrain frequent pattern tree and its construction method [J]. Journal of Chinese Computer Systems, 2010, 31(4): 682 – 685. (in Chinese)

[16] Unil Yun, Keun Ho Ryu. Approximate weighted frequent pattern mining with/without noisy environments [J]. Knowledge-Based Systems, 2011, 24(1): 73 – 82.

[17] 宋国杰, 唐世渭, 杨冬青, 等. 数据流中异常模式的提取与趋势监测 [J]. 计算机研究与发展, 2004, 41(10): 1754 – 1759.

Song Guojie, Tang Shiwei, Yang Dongqing, et al. Extraction and trend detection of unusual patterns over data streams [J]. Journal of Computer Research and Development, 2004, 41(10): 1754 – 1759. (in Chinese)

[18] Kamber M, Han J, Jenny Y C. Metarule-guided mining of multi-dimensional association rules using data cubes [C] // Proceeding of Knowledge Discovery and Data Mining, 1997: 207 – 210.

[19] Ananthanarayana V S, Narasimha Murty M, Subramaninan D K. Multi-dimensional semantic clustering of large databases for association rule mining [J]. Pattern Recognition, 2001, 34(4): 939 – 941.

[20] Li Q, Feng L, Wong A. From intra-transaction to generalized inter-transaction: Landscaping multidimensional contexts in association rule mining [J]. Information Science, 2005, 172(3/4): 361 – 395.

[21] Kweku M, Osei B. A context-aware data mining process model based framework for supporting evaluation of data mining results [J]. Expert Systems with Applications, 2012, 39(1): 1156 – 1164.

[22] Aggarwal C C, Han J, Wang J, et al. A framework for on-demand classification of evolving data streams [J]. IEEE Transactions On Knowledge and Data Engineering, 2006, 18(5): 577 – 589.

[23] Dong G, Li J. Efficient mining of emerging patterns: Discovering trends and differences [C] // Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD-99), 1999: 43 – 45.

[24] Song H S, Kim J K, Kim S H. Mining the change of customer behavior in an internet shopping mall [J]. Expert Systems with Applications, 2001, 21(3): 157 – 168.

[25] Chen M C, Chiu A L, Chang H H. Mining changes in customer behavior in retail marketing [J]. Expert Systems with Applications, 2005, 28(4): 773 – 781.

Customer behavior pattern mining and change detection with context intensity constraints

JU Chun-hua[1,2], *SHUAI Zhao-qian*[1]

1. Information College, Zhejiang Gongshang University, Hangzhou 310018, China;
2. Center for Studies of Modern Business, Zhejiang Gongshang University, Hangzhou 310018, China

Abstract: Currently, many studies dedicated to context aware based recommendation, considered different types of context properties, but they ignore the important degree of different context attribute impact the behavior, that is, context strength. This paper defines the customer context, context intensity, and behavior changing quantitatively; presents context strength constrained pattern mining methods and change detecting method to extract the critical situation caused by changes in behavior. The proposed algorithm increased the sensitivity of the interests changing, improvement of the massive data under support for sparse association rules, the shortcomings of low confidence sensitivity. Experiments and analysis demonstrated the feasibility and effectiveness.

Key words: context intensity; customer behavior; constraint-based frequent patterns; interests drift detecting; recommended strategy

附表 1摇打标签后含情境知识的购物篮数据片段（前0/1 000）

T_UID	CID	CPI	天气	季节	节假日	性别	婚姻	学历	收入	城市	商品类目 1	商品类目 2	商品类目 3	商品类目 4
1	1	CPI：[-∞,-3%）	舒适	春	节假日	男	已婚有孩子	研究生及以上	高	一线城市	鞋、箱包	鞋、箱包	鞋、箱包	男装
2	2	CPI：[-∞,-3%）	舒适	夏	非节假日	女	已婚有孩子	研究生及以上	高	一线城市	运动、户外	男装	化妆品	运动、户外
3	3	CPI：[-∞,-3%）	持续高温	秋	非节假日	男	已婚有孩子	研究生及以上	高	一线城市	男装	家用电器	运动、户外	鞋、箱包
4	4	CPI：[-∞,-3%）	舒适	冬	重要节假日	女	已婚无孩子	研究生及以上	高	中小城市	女装	运动、户外	鞋、箱包	鞋、箱包
5	5	CPI：[-∞,-3%）	降温	冬	重要节假日	女	已婚无孩子	研究生及以上	中	一线城市	男装	鞋、箱包	鞋、箱包	化妆品
6	6	CPI：[-∞,-3%）	持续高温	秋	非节假日	男	年轻单身	研究生及以上	高	中小城市	化妆品	女装	运动、户外	男装
7	7	CPI：[-∞,-3%）	持续低温	春	非节假日	男	年轻单身	研究生及以上	高	中小城市	男装	运动、户外	女装	女装
8	8	CPI：[-∞,-3%）	升温	夏	非节假日	女	已婚有孩子	研究生及以上	高	一线城市	运动、户外	母婴产品码	家用电器	运动、户外
9	9	CPI：[-∞,-3%）	舒适	秋	非节假日	男	年轻单身	研究生及以上	中	一线城市	女装	女装	化妆品	鞋、箱包
10	10	CPI：[-∞,-3%）	舒适	冬	非节假日	女	年轻单身	研究生及以上	中	一线城市	手机、数码	女装	女装	鞋、箱包
11	11	CPI：[-∞,-3%）	持续高温	夏	非节假日	男	已婚有孩子	本科及以下	高	一线城市	女装	化妆品	运动、户外	男装
12	12	CPI：[-∞,-3%）	舒适	秋	重要节假日	女	已婚有孩子	本科及以下	高	一线城市	运动、户外	鞋、箱包	家用电器	女装
13	13	CPI：[-3%,3%）	舒适	春	非节假日	男	已婚有孩子	本科及以下	中	中小城市	男装	女装	化妆品	女装
14	14	CPI：[-3%,3%）	升温	夏	非节假日	女	已婚有孩子	本科及以下	中	中小城市	手机、数码	女装	男装	运动、户外
15	15	CPI：[-3%,3%）	舒适	秋	非节假日	女	已婚无孩子	本科及以下	高	一线城市	运动、户外	女装	化妆品	家用电器
985	10	CPI：[-3%,3%）	降温	春	非节假日	男	年轻单身	研究生及以上	中	一线城市	男装	女装	男装	女装
986	11	CPI：[-3%,3%）	降温	夏	非节假日	女	已婚有孩子	本科及以下	高	一线城市	女装	鞋、箱包	女装	鞋、箱包
987	12	CPI：[-3%,3%）	降温	秋	非节假日	男	已婚有孩子	本科及以下	高	一线城市	母婴产品	男装	家用电器	鞋、箱包
988	13	CPI：[-3%,3%）	升温	冬	节假日	女	已婚有孩子	本科及以下	中	中小城市	女装	手机、数码	化妆品	鞋、箱包
989	14	CPI：[-3%,3%）	持续低温	秋	非节假日	男	已婚有孩子	本科及以下	中	中小城市	化妆品	女装	鞋、箱包	鞋、箱包
265	15	CPI：[-3%,3%）	持续高温	冬	节假日	男	已婚无孩子	本科及以下	高	一线城市	运动、户外	男装	运动、户外	男装
266	16	CPI：[-3%,3%）	持续低温	秋	非节假日	女	已婚无孩子	本科及以下	高	一线城市	化妆品	家用电器	鞋、箱包	男装
267	17	CPI：[3%,1%）	持续低温	夏	非节假日	男	已婚无孩子	本科及以下	中	一线城市	男装	运动、户外	鞋、箱包	鞋、箱包
268	18	CPI：[3%,1%）	降温	秋	非节假日	女	已婚无孩子	本科及以下	中	中小城市	女装	女装	家用电器	鞋、箱包
269	19	CPI：[-∞,-3%）	持续高温	夏	非节假日	男	已婚有孩子	本科及以下	中	中小城市	母婴产品	母婴产品	家用电器	运动、户外
270	20	CPI：[-∞,-3%）	降温	秋	非节假日	女	年轻单身	本科及以下	高	一线城市	女装	女装	运动、户外	化妆品
271	21	CPI：[-∞,-3%）	持续高温	冬	非节假日	女	年轻单身	本科及以下	高	一线城市	男装	家用电器	化妆品	运动、户外
305	17	CPI [3%,+∞）	升温	夏	非节假日	女	已婚无孩子	本科及以下	中	一线城市	手机、数码	手机、数码	化妆品	男装
306	18	CPI：[-∞,-3%）	持续高温	秋	重要节假日	男	已婚无孩子	本科及以下	中	中小城市	手机、数码	运动、户外	女装	家用电器
307	28	CPI：[-3%,3%）	舒适	春	非节假日	男	年轻单身	本科及以下	中	一线城市	男装	家用电器	鞋、箱包	鞋、箱包